1 Data, Data Management, and Reproducible Research in Linguistics: On the Need for *The Open Handbook of Linguistic Data Management*

Andrea L. Berez-Kroeker, Bradley McDonnell, Lauren B. Collister, and Eve Koller

> Data have no value or meaning in isolation; they exist within a knowledge infrastructure—an ecology of people, practices, technologies, institutions, material objects, and relationships.
> —Borgman, *Big Data, Little Data, No Data*

1 Introduction: Why a "handbook" on data management in linguistics?

Data, in many forms and from many sources, underlie the discipline of linguistics. We feel it would not be hyperbolic to say that data are the lifeblood of research, and proper management of data collections is essential to the future of our field. From descriptive to theoretical work, from corpus-based to introspection-based inquiry, from quantitative to qualitative analysis, linguists rely on data every day. Although technologies for producing, managing, and analyzing vast amounts of data have developed in recent decades, we must recognize that linguists have long depended on data to develop generalizations and theories about the nature of human language, even since the field's earliest forays into philology. We see the potential of linguistic data to inform the field and inspire future scientific inquiry and innovation into the nature of humanity through language. To unlock this potential, data must be understandable, discoverable, reusable, shareable, remixable, and transformable. All data sets, from inscriptions on stone tablets to introspective grammaticality judgments to terabytes of recordings of sociolinguistic interviews, must be managed conscientiously and carefully. There is no doubt that managing data requires time, effort, and, in many cases, specialized training, and, as long as technology allows us to collect and use ever-increasing amounts of data, it likely always will.

In developing this Handbook, we followed Borgman's (2015:29) definition of *data* as "entities used as evidence of phenomena for the purposes of research or scholarship." This definition is particularly apt for the field of linguistics, with its many and diverse data sources and forms. We also know that data do not exist in a vacuum, as the epigraph to this chapter indicates. Data, especially in linguistics, are a representation of the people who provided them, so we need to take care to manage data in ethical ways that respect the dignity and autonomy of everyone involved. In this way, the same linguistic data can also serve humanistic endeavors.

Data management is far more than just storing data. It entails a broad range of tasks, as data need to also be collected, cataloged, organized, annotated, described, processed, analyzed, preserved, shared, and cited, if linguistics is to be an open and transparent social science. Implementing these procedures on a broad scale ultimately enables *reproducible research*, which the National Academies of Sciences, Engineering, and Medicine define as research that obtains "consistent results using the same input data; computational steps, methods, and code; and conditions of analysis" (National Academies of Sciences, Engineering, and Medicine 2019:6–7). The upshot of taking this approach is providing greater scientific accountability through facilitating access for other researchers to the data upon which research conclusions are based (see, e.g., Buckheit & Donoho 1995; de Leeuw 2001; Donoho 2010; Berez-Kroeker et al. 2018; see also Gawne & Styles, chapter 2, this volume, for more about reproducible research).

Historically, methods for managing data in our field have been developed somewhat in isolation. Different subfields, research labs, and even individual researchers have developed their own practices and expectations regarding proper management of data. Even though everyone who uses data must manage them at some level, the isolated development of data management methods has meant that there has been very little discussion across

the discipline concerning any commonalities or shared best practices that might exist.

Furthermore, despite the time and care that goes into proper data management for the service of reproducible research, our field has only cursorily acknowledged the scholarly value of this work, electing instead to prioritize analysis and theory in publications. Reflecting deeply, either in writing or in practice, on one's data management practices as "meta-research" has been relegated to a second class of publication and scholarship, less overtly valued than a new research paper or book, and has not been widely encouraged outside of the development of specifically instructional textbooks. In fact, data management has often been thought of as an afterthought or as "somebody else's job," with the assumption that librarians or data specialists are the ones who will care for this work (Mons 2018:27).[1] While institutional support in maintaining data is crucial for its discovery and long-term survival, the researchers who collect the data are an essential part of the ongoing care and description of data, and their conscious involvement early on make these data sets usable for future scholars (Borgman 2015:275). Data work should be valued as highly as any other aspect of research.

Fortunately, the field is changing. We have begun to acknowledge the time, care, and expertise that go into proper data management in service of reproducible linguistics. Whereas previously the primary outputs of research were publications almost exclusively in the realm of theory and analysis, it is now increasingly common to see the data themselves as an important output, worthy of valuation in hiring, tenure, and promotion (see Alperin et al., chapter 13, this volume). For example, in 2018 the Linguistic Society of America adopted the Statement on the Evaluation of Language Documentation in Hiring, Tenure, and Promotion,[2] which gives suggestions for evaluating data and other non-traditional research outputs. Furthermore, it is not uncommon now to reflect on one's data management practices in writing, thus creating transparency about data sources and research methods.

Linguists are starting to confront data management issues and make visible the need for better data practices. More than 25 years ago, the issue of questionable data management practices compelled Sally Thomason, the then-editor of *Language*, a top journal in the field, to write a five-page column about her observations regarding the data behind articles submitted to the journal (Thomason 1994; see also Thomason, foreword, this volume). In it, she describes her realization that verifying all data for accuracy is too cumbersome a task to fall solely on the shoulders of the journal editor, and how, being human, she needed to "rely on the assumption that the data in accepted papers is generally correct." Nonetheless, she continues,

> Because of the traditionally high standards of *Language* regarding linguistic data, I have tried to identify cases where I may need to pay special attention to the accuracy of data: cases where the referees found problems with the data, where the data seems to be incompletely attested, or where a spot check reveals errors. When I began my term as editor, I expected that there would be cases of this kind from time to time. I did not expect that these cases would occur frequently—so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable. (409)

Thomason further notes that her concerns about data did not resonate particularly loudly. She provides as an example an interaction with an author in whose submission she found numerous small problems with the data (e.g., incorrect morphological parsing, incorrect glossing): "the author's response on being informed of the errors was disturbing. The most serious mistakes, I was told, came from a theoretical article that the author had cited extensively; it was the authors of that article who were responsible for the mistakes, not the author of the *Language* paper" (410).

Clearly, basing linguistic theory on inaccurate data is harmful for theory, especially when the theory builders are willing to look the other way when confronted with problematic data. Thomason compares the author's reaction to a "What's wrong with this picture?" children's puzzle:

> What you see in such a puzzle is a scene that looks perfectly normal at first glance, but that on closer inspection turns out to have impossible features, like a man sitting comfortably in a chair that has only one leg. A linguistic theory that rests on false or inadequate data is like the man in the chair with one leg: the support is illusory. (410)

Thomason then goes on to offer advice to authors regarding proper handling of the data that underlies publications. Even today her advice still rings true and has led to several recent initiatives. Beginning in the mid-2010s, researchers began to examine *transparency* in linguistics research. Transparency involves making details about research practices explicit in publications: some examples are whether data were collected directly by the author or taken from a published or archival

source, whether data were collected in a lab or in the field, what hardware and software tools were used to collect and process data, presenting demographic information about language consultants or interviewees. Gawne et al. (2017) surveyed one hundred descriptive grammars for how forthcoming authors were about their data and collection methods; Schembri (2019) and Hochgesang (2019) both surveyed methodological transparency in sign language linguistics; Gawne et al. (2019) examines transparency of authors in gesture studies; Berez-Kroeker, Gawne et al. (2017) look at transparency in articles from nine journals across the discipline over a ten-year span. In all cases, it was found that authors' attention to conveying details about data and data collection to readers was lacking in some respect (see Berez-Kroeker et al. 2018 for a discussion). This includes, in particular, citing excerpts of data (e.g., interlinearized glossed text or other numbered examples) back to their sources in a way that makes them retrievable by a reader, especially when the data come from a source other than a traditional paper publication, such as a book or a journal article.

Also in the mid-2010s, more than forty linguists and data specialists participated in a multiyear project to develop standards and recommendations for one particular aspect of data management, the citation of data sets in linguistics publications.[3] This group produced a position paper identifying barriers to better citation and attribution of data as a part of linguistic research (Berez-Kroeker et al. 2018). Over time the group evolved into the 100+-member Linguistic Data Interest Group,[4] formed as part of the much larger Research Data Alliance,[5] an international organization that at the time of writing has over nine thousand members from 137 countries. Among the outputs of the project and the Research Data Alliance group are a guide to help linguists understand the value of data citation, known as *The Austin Principles of Data Citation in Linguistics* (Berez-Kroeker, Andreassen et al. 2017),[6] and a set of standardized formats for citing data sets in publications, known as the *Tromsø Recommendations for Citation of Research Data in Linguistics* (Andreassen et al., 2019; see Conzett & De Smedt, chapter 11, this volume).

2 About this Handbook

The editors and many authors in this Handbook have been and continue to be active participants in the Research Data Alliance Linguistic Data Interest Group. Our work with this group has made it clear that even 25 years after Thomason's editorial column, the discipline of linguistics still does not have a culture of broad and open discussion about data. Many linguists we have consulted with have lamented the lack of academic reward for data work, and still more have admitted that they simply do not know very much about how to manage their data in a way that ensures they will be citable, accurate, shareable, and sustainable, nor do they know much about their colleagues' practices in these areas. Despite the barriers, however, the reality of linguistic practice today is that most of us use data, most of us wish to use them thoroughly and carefully, many of us share data and code with our colleagues, and most of us have some methods for managing data, whether or not those methods have been codified. Thus, this Handbook grew out of a need to provide a forum in which researchers could share their data management practices with the aim of learning more about the current state of data work across the field. In this way, we hope that the discipline can reflect deeply about the past and present and foster an open conversation about the future of data work in linguistics.

This Handbook is divided into two parts. Each of the full-length chapters in part I delves into prominent issues surrounding data and data management. Part II consists of shorter *data management use cases*, each of which demonstrates a concrete application of the abstract principles of data management in specific studies, some actual and some hypothetical.

2.1 Part I overview

Because of the nature of linguistic study, creating reproducible data involves considerations such as the ethics of working with and for the benefit of human participants, copyright over creative works and expressions, and techniques for transforming data. These considerations come into play throughout the data management process, from the conception of the study to the development of mid-project file-naming conventions to the final steps of archiving, sharing, and tracking the use of data.

Chapters 2–4 provide details on important conceptual foundations of data management for linguistic data. In chapter 2, Lauren Gawne and Suzy Styles discuss the place of linguistics in the broader data movements across the social sciences, in which openness of data and

transparency of research methodologies have become a central concern. In chapter 3, Jeff Good provides a comprehensive survey of the diversity of data types that are used within linguistics, including data that directly represent observable linguistic behavior as well as secondary data types used to support linguistic analysis. In chapter 4, Gary Holton, Wesley Y. Leonard, and Peter L. Pulsifer discuss the range of ethical considerations that are necessary when working with linguistic data of all kinds.

Chapters 5 and 6 present important principles in the implementation of data management. In particular, in chapter 5, Eleanor Mattern covers the life cycle of data, emphasizing consistency and a future-minded orientation while outlining best practices for creating sustainable, reusable, long-lasting data. In chapter 6, Na-Rae Han discusses the processes behind transforming raw data into usable data, including transliteration, loss, augmentation, and corruption of information, all while maintaining reversibility through judicious employment of version control.

Chapters 7, 8, and 9 provide helpful guidance to researchers planning to collect, manage, and share their own data. In chapter 7, Helene N. Andreassen gives advice on archiving one's research data, from preparing data for archiving to selecting an appropriate repository that ensures long-term preservation, retrieval, and visibility. In chapter 8, Susan Smythe Kung describes the process of developing a data management plan, a document that allows researchers to clearly articulate plans for data collection, processing, preservation, and sharing; data management plans are often required by funding organizations and can save researchers time, money, and frustration. In chapter 9, Lauren B. Collister provides a foundation for understanding copyright and its interaction with data, sharing practices to implement throughout the data life cycle to reduce legal barriers to the open sharing and publication of linguistic data.

Chapters 10 and 11 are about finding and reusing prepared linguistic data sets. In chapter 10, Laura Buszard-Welcher discusses how we can move human knowledge into the future by preserving linguistic data—both from living languages and from archived data from sleeping languages—for the long term. In chapter 11, Philipp Conzett and Koenraad De Smedt provide concrete guidance on how to properly cite data through bibliographic references and in-text citations.

Finally, chapters 12 and 13 discuss the valuation of the time and effort involved in data management as a research endeavor. In chapter 12, Robin Champieux and Heather L. Coates describe metrics that can reveal the usage of a data set and help tell the story of its impact beyond the initial research study. In chapter 13, Juan Pablo Alperin, Lesley A. Schimanski, Michelle La, Meredith T. Niles, and Erin C. McKiernan follow with an analysis of the role of data in review, promotion, and tenure based on their extensive corpus of documents.

2.2 Part II overview

The second part of the Handbook provides snapshots of current practices in the form of data management use cases. These come from a sampling of subfields and represent only a selection of the many data management practices available in the field. These use cases were selected to include both signed and spoken languages, and collectively they cover vast swaths of linguistic research, including sociolinguistics, discourse and conversation analysis, language documentation and description, language reclamation, historical linguistics and language change, first- and second-language acquisition, computational applications such as forced alignment and speech recognition, corpus linguistics, experimental linguistics, syntax, psycholinguistics, neurolinguistics, phonology, typology, and semantics. Many more use cases could have been included, and in an effort to make our data management practices and research methodologies more transparent, we editors encourage readers to start a practice of writing your own use cases.[7]

2.3 How to use this Handbook

The aim of the Handbook is to provide a snapshot into current practices, some of which are well established, while others are more cutting-edge. As this is, to our knowledge, the first handbook of its kind, it is not meant to serve as a comprehensive manual or textbook. Nonetheless, there are ways to incorporate the principles and practices described herein into one's own practices, research program, and/or career. We expect the Handbook to be valuable to a broad audience, including students, early career and seasoned researchers, and instructors at many levels. The Handbook can be used as a primary resource for classroom courses on linguistic data management, or selected chapters can serve as examples for management methods in courses focused on particular subfields.

Self-study is also possible, and we have developed an online open access companion course for this

volume, available to anyone, free of charge, at http://linguisticdatamanagement.org.[8] The online course contains synopses of each part I chapter, including keywords to learn, activities to reinforce principles of data management, suggestions for implementing better data management in one's career, quizzes to test your understanding, and cross-referenced links to relevant data management use cases.

3 Acknowledgments

We wish to offer our sincere gratitude to many people who made this Handbook possible. First, we thank the more than one hundred authors who willingly shared their knowledge and experience with our readers. Importantly, because this volume has a pan-linguistics scope, we needed to rely on many experts across the discipline to help us understand the "data scene" in different subfields, as well as identify appropriate authors to contribute. These "area advisors" included Claire Bowern, Kathleen Currie Hall, Na-Rae Han, Heidi B. Harley, Kristine Hildebrandt, Julie Hochgesang, Barb Kelly, Tyler Kendall, Emma Marsden, Steven Moran, and Luca Onnis. Emily Bender, Colleen Fitzgerald, Helen Aristar-Dry, and Eric Bakovic provided valuable discussion on the volume. Special thanks to Susan Smythe Kung, Gary Holton, Peter L. Pulsifer, and the participants in the Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics project.

Much appreciation goes to Heidi B. Harley and the entire editorial board of the Open Handbooks in Linguistics initiative for envisioning the series and giving us the opportunity to develop one of its first publications; we also thank Philip Laughlin, Alex Hoopes, Marc Lowenthal, and Anthony Zannino at MIT Press for their editorial support. Allison Silver Adelman of Silver Academic Editing provided professional assistance with the substantial task of consistently formatting fifty-five chapters and developing the index. Eve Koller, Shirley Gabber, and Dannii Yarbrough together developed the online companion course.

The editors would like to acknowledge the University Library System at the University of Pittsburgh for their support of Lauren B. Collister's editorial work on this volume, as well as their open access journal publishing services, which housed the platform for the management of chapter submission and reviews as part of their Scholarly Exchange service. We would also like to thank the Public Knowledge Project and the many volunteers who work on Open Journal Systems, which made our jobs as editors much easier and better organized. We also acknowledge the support of the Department of Linguistics and the College of Languages, Linguistics, and Literature at the University of Hawai'i at Mānoa and the Faculty of Culture, Language, and Performing Arts at Brigham Young University Hawai'i.

Finally, this material is based on work supported by the National Science Foundation under grants SMA-1649622 and SMA-1745249. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Notes

1. We would like to emphasize that we reject the dual implication that data work is not part of linguistics research and that the contributions of librarians in describing, cataloging, preserving, and sharing research is somehow less valuable than that of researchers.

2. https://www.linguisticsociety.org/sites/default/files/Evaluation_Lg_Documentation.pdf.

3. This project was called Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics (NSF SMA-1447886). For details on activities see https://sites.google.com/a/hawaii.edu/data-citation/.

4. https://rd-alliance.org/groups/linguistics-data-ig.

5. https://rd-alliance.org/.

6. The Austin Principles (http://site.uit.no/linguisticsdatacitation/) are essentially the FORCE11 *Joint Declaration of Data Citation Principles* (Martone 2014), annotated for linguistics.

7. If you would like to share your own linguistic data management use case, please upload it to our collection on Zenodo: https://zenodo.org/communities/ldmuc/.

8. For readers who are reading this in the print version, the Handbook itself is also available online, open access, and free of charge.

References

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. *Research Data Alliance*. https://doi.org/10.15497/rda00040.

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, and the Linguistics Data Interest Group. 2017. *The Austin Principles of Data Citation in Linguistics*. Version 1.0. http://site.uit.no/linguisticsdatacitation/austinprinciples/. Accessed March 17, 2020.

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly, and Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003–2012. https://sites.google.com/a/hawaii.edu/data-citation/survey. Accessed March 17, 2020.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56:1–18. https://doi.org/10.1515/ling-2017-0032.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Buckheit, Jonathan B., and David L. Donoho. 1995. WaveLab and reproducible research. In *Wavelets and Statistics*, ed. Anestis Antoniadis and Georges Oppenheim, 55–81. New York: Springer.

De Leeuw, Jan. 2001. Reproducible research: The bottom line. *UCLA Department of Statistics Papers*. http://escholarship.org/uc/item/9050x4r4. Accessed March 17, 2020.

Donoho, David L. 2010. An invitation to reproducible computational research. *Biostatistics* 11:385–388.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, and Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation and Conservation* 11:157–189.

Gawne, Lauren, Chelsea Krajcik, Helene N. Andreassen, Andrea L. Berez-Kroeker, and Barbara F. Kelly. 2019. Data transparency and citation in *Gesture*. *Gesture*.

Hochgesang, Julie A. 2019. Sign language description: A Deaf retrospective and application of best practices from language documentation. Paper presented at the Signed and Spoken Language Linguistics Conference, Osaka.

Martone, M., ed. 2014. *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles*. San Diego: FORCE11. https://doi.org/10.25490/a97f-egyk, https://www.force11.org/group/joint-declaration-data-citation-principles-final.

Mons, Barend. 2018. *Data Stewardship for Open Science: Implementing FAIR Principles*. New York: Taylor and Francis. https://doi.org/10.1201/9781315380711.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science.

Schembri, Adam. 2019. Making visual languages visible: Data and methods transparency in sign language linguistics. Paper presented at Theoretical Issues in Sign Language Research Conference, Hamburg, September 26–28.

Thomason, Sarah. 1994. The editor's department. *Language* 70 (2): 409–413.

This is a section of doi:10.7551/mitpress/12200.001.0001

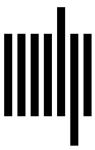
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363